DETECTION OF SPEECH LANDMARKS USING TEMPORAL CUES

Ariel Salomon and Carol Espy-Wilson

Electrical and Computer Engineering Department, Boston University
8 St. Mary’s St., Boston, MA 02215
{ariel,espy}@lexicon.bu.edu

Abstract

In order to improve the performance of speech recognizers, particularly in degraded environments, it may be beneficial to integrate use of temporal information. As literature has shown that human listeners are able to use temporal cues in speech recognition tasks, this study examines algorithms for extraction of temporal cues in a speech signal. The task under analysis is the location of landmarks to direct further analysis in a knowledge-based speech recognition system. In this study, a set of robust landmark types were located with an accuracy of 82.5%–100.0%.

1 Introduction

A major problem in the development of speech recognition systems is the understanding of speech in noise, or from reduced spectral information. The study of speech perception in noise has a long history, and the particular spectral contributions to speech intelligibility are well understood, note for example the work of Miller & Nicely [1]. However, recent studies show that another source of information that may be of use [2], particularly in degraded environments [3], consists of temporal cues, i.e. the temporal structure of components of the speech signal. These cues are not targeted by traditional speech recognition systems, which generally focus on spectral features, using data-derived spectral templates (but note also the approach of Sharma & Hermansky [4]). The auditory system is sensitive to subtle temporal effects, such as phase locking of auditory nerve firing to periodic signals, which suggests that this information is available for use in human speech recognition. This is shown to be the case in studies by Van Tasell et al. [5], Shannon et al. [6], and Turner et al. [7] which demonstrate the ability of human listeners to recognize speech—particularly consonant manner, nasality, and voicing—from primarily temporal cues.

In this study, we define temporal information as information present in the structure of envelopes of bandpass components of the speech signal. A productive system of categorization for temporal information has been suggested by Rosen [8], who proposes three categories of temporal information in speech: (1) “envelope information” (with fluctuations at rates from 2 to 50Hz) which contains amplitude and duration cues to manner of articulation and voicing, as well as information about vowel identity (e.g., vowel length) and prosodic cues; (2) “periodicity information” (fluctuations at rates from approximately 50 to 500Hz) provides cues to voicing which can aid in manner identification, as well as marking stress locations by changes in pitch; and (3) “fine structure” (fluctuations at higher rates) which provides information about spectral shape which is most useful for identifying place of articulation for consonants and vowel quality, though there are some cues for manner such as the high-frequency content of many obstruents. These imply that use of temporal information for a recognizer (particularly when combined with spectral cues) should concentrate on two major types of information: low-frequency envelope, and periodicity associated with pitch.

The goal of this work is the development of a set of algorithms which can be integrated into the front end of a knowledge-based speech recognition system. The algorithms under discussion locate acoustically abrupt events in a speech signal, toward the discovery of landmarks in speech, or linguistically important points of analysis in the signal. Landmarks are a phonologically motivated set of locations in the signal at which further feature extraction is required, for the purpose of classifying adjacent or surrounding regions. Locating a set of landmarks in the speech signal is intended to direct an efficient analysis of linguistically-relevant details in the signal.

2 Features

Front-end processing to generate envelopes uses a 60-channel auditory gamma-tone filterbank with characteristic frequencies (CFs) as per Carney [9, 10] and ranging from 100 to 8000Hz. The filter bank is followed by a Hilbert transform envelope filter.

The major parameters used for event location in this project are from an onset/offset detector based on a first difference measure, with per-channel computation as in (1):

\[
D_{i,k}(n) = 20 \log \sum_{m=-\infty}^{\infty} x_i(n + m)w_1(m) \quad (1)
\]

ISCA Archive
http://www.isca-speech.org/archive
$$-20 \log \sum_{m=-\infty}^{\infty} x_i(n + m - k)w_2(m - k)$$

From this, two summary measures are computed: channels with positive differences (increasing magnitude) are summed to produce an 'onset' signal, and channels with negative differences (decreasing magnitude) are summed to produce an 'offset' signal, both in units of decibels (dB):

$$o(n) = \frac{1}{N} \sum_{i \in \{k: D_{i,k}(n) > 0\}} D_{i,k}(n) \quad (2)$$

$$o(n) = \frac{1}{N} \sum_{i \in \{k: D_{i,k}(n) < 0\}} D_{i,k}(n) \quad (3)$$

The low-frequency amplitude envelope fluctuations (or more specifically, the sharpest changes in this envelope) are captured in this work by the peaks in the energy difference parameters. To capture the midrange periodicity component of temporal information, a pitch detection algorithm was included. An estimate of pitch is computed by producing estimates of the fundamental period in individual channels using the Average Magnitude Difference Function and combining estimates by way of a modified histogram (sum of confidence measures at each sample delay), as shown in Figure 1. Improved response of the energy difference parameters were obtained by adjusting the difference length parameter $k$ as a function of periodicity and fundamental period: decreasing it when a channel is silent to sharpen response to onsets from silence (e.g. stop bursts), increasing it in aperiodic regions, and tuning to twice the pitch period in periodic regions.

Figure 2: Parameter extraction results for the sentence “The saw is broken.” a) spectrogram; b) raw summary periodicity and aperiodicity (marked with ‘+’) confidence scores; c) smoothed pitch estimate in periodic regions, aperiodicity confidence ‘(−)’ in aperiodic regions; d) onset and offset parameters, chosen peaks labeled with stems; e) detected landmarks. Labels from TIMIT transcription are superimposed, both the words (top) and phonemes (bottom). Vertical lines mark labeled phoneme boundaries in (a)-(d), landmark locations in (e).

3 Landmark extraction

For the purpose of detection, a set of landmark types based on acoustic parameters are defined, as listed in Table 1. An example set of extracted landmarks is shown as Figure 2(e). These correspond to the polarity (onset or offset) and correlation with periodicity content—either at a boundary where periodic content begins or ends (correlated with voicing onset or offset), surrounded by periodic excitation (correlated with sonorant consonants), or outside of periodic excitation and/or correlated with a boundary of aperiodic excitation (correlated with obstruct consonants). These are the types of events output by the event detector.

The landmark generation process begins by locat-
ing regions that contain periodic excitation and aperiodic excitation, based on thresholds of the corresponding confidence metrics (and some pruning, combining). An example of this process is illustrated in Figure 2, in the transition from (b) to (c). Peaks are located in the onset and offset measures using thresholds on height of the peaks and a minimum dip between peaks, as shown in Figure 2(d).

These two types of information are combined to locate landmarks. Onset/offset events near the beginning/end of a periodic region are labeled as $+v$/$-v$. Correspondingly, onset/offset events near the beginning/end of an aperiodic region are labeled as $+c$/$-c$. The criterion for this locality is determined by a set of thresholds which have been trained. Remaining boundaries of these regions are labeled as landmarks of the same types but the times are less accurate. Remaining onset/offset events are labeled as $+c$/$-c$ if they are outside of a periodic region, or $+s$/$-s$ within a periodic region.

A set of expected landmarks must be generated from the phoneme-labeled transcriptions available in the TIMIT database, for the purpose of comparing with the reference transcription. These are generated using a simple rule-based algorithm based on manner class of adjacent segments, and are known to have some inherent error at this level of representation. Some of this error is accounted for by inserting events which are considered ‘nonrequired’ because they are possible, and should be caught by the matching algorithm, but are not necessarily strongly expected. For example, it is common to find an offset following a strong stop burst, often at the boundary between friction and aspiration; also, there will often be a double burst for a velar stop—however, it is not unlikely in either case that the event will not be present.

### Table 1: Landmark types

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Description (examples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$+v$</td>
<td>voicing onset</td>
<td>onset corresponding to beginning of periodicity (beginning of a vowel or sonorant consonant)</td>
</tr>
<tr>
<td>$-v$</td>
<td>voicing offset</td>
<td>offset corresponding to end of periodicity (end of a vowel or sonorant consonant)</td>
</tr>
<tr>
<td>$+s$</td>
<td>sonorant onset</td>
<td>onset within periodic region (onset at release of nasal or semivowel)</td>
</tr>
<tr>
<td>$-s$</td>
<td>sonorant offset</td>
<td>offset within periodic region (offset at closure for nasal or semivowel)</td>
</tr>
<tr>
<td>$+c$</td>
<td>obstruent onset</td>
<td>onset corresponding to beginning of aperiodicity (stop consonant burst, affricate or fricative onset)</td>
</tr>
<tr>
<td>$-c$</td>
<td>obstruent offset</td>
<td>offset corresponding to end of aperiodicity (stop, affricate or fricative offset)</td>
</tr>
</tbody>
</table>

### 4 Results

Training was performed using a set of 20 sentences (spoken by 10 males, 10 females) randomly drawn from the TIMIT training set. Testing was performed using all 120 sxx (phonologically compact) sentences from a subset of the TIMIT test set referred to as the TIMIT core test set, which contains a well-balanced and complete set of phonological contexts.

At this point, the system works reasonably well for extracting events that are strongly present (and, for the purpose of comparison, consistently labeled). Some adjustment and training has been performed on a number of the time and energy thresholds involved in the system, in particular including those discussed in Section 3.

Performance of the pitch detector was compared with the Entropic ESPS `get_f0` component as a reference. The voicing decisions agreed 89.1% of the time on the training set (88.7% on the test set). Error in pitch decisions, scaled relative to median pitch for each utterance, was 12.0% (13.4% test). It was noted that a major type of error was that both pitch detectors occasionally chose a pitch of one half the correct pitch (twice the correct pitch period). This is an understandable type of error due to the time-domain peak picking that is used in the temporal detector, and is more likely in female speech because the male pitch period lengths are often above one half the 20ms window length used (if pitch $\leq 100$Hz). In order to determine how often this occurred, frames were counted in which the error between the two detectors was improved by multiplying one or the other by a factor of 2. For the test set, this occurred in 0.55% (0.91% for female speech) of the frames for the `get_f0` detector, and in 1.62% (2.78% for female speech) of the frames.
Table 2: Major sources of error (% of misses)

<table>
<thead>
<tr>
<th>Source</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>vowel-sonorant boundaries</td>
<td>42.0%</td>
<td>39.9%</td>
</tr>
<tr>
<td>weak fricative boundaries</td>
<td>20.8%</td>
<td>23.4%</td>
</tr>
</tbody>
</table>

for the temporal detector. Performing the adjustments reduced pitch error to 4.91% (8.58% on the test set, concentrated in the male speech).

Landmarks were detected with overall detection rate of 75.4% (70.0% on the test set), with an insertion rate of 9.8% (11.6%). Nearly half of the error rate was due to missed boundaries between sonorant consonants and vowels, an event type that was detected with only 46.7% (41.6%) accuracy. Even in the transcription of the TIMIT database [11], a rule was used to insert these boundaries because they were not robustly locatable. Not counting the two major sources of error (see Table 2), the detection rate was 86.5% (82.6%). More detailed results for strongly expected event types are listed in Table 3.

5 Conclusion

This work shows that use of temporal information for landmark detection is feasible, and this implementation performs reasonably well. However, the implementation has a number of key areas where it could be improved, particularly use of prediction and/or long term integration of information (especially in the pitch detector).

It is also the case that higher level processing needs to be developed to generate a segment or phone level representation. Further interpretation of landmarks would need to add use of duration cues for full use of temporal information.

6 Acknowledgements

This work was supported by NSF grant #SBR-9729688.

References


